Data Science at Work: Linking Unmapped and UNASSIGNED Structured-Abstract Labels to the Five Canonical NLM Categories

December 1, 2021

François-Michel Lang

1 Background

The 2021 Baseline contains 4,116,876 MEDLINE citations with Structured Abstracts (SAs) (15% of the 27,321,214 MEDLINE citations in the 2021 Baseline) which in turn contain 17,418,153 instances of labels. 14,890,069 (85.49%) of these label instances are well constructed in the sense that they are linked to one of the five canonical NlmCategories (BACKGROUND, OBJECTIVE, METHODS, RESULTS, and CONCLUSIONS). A few examples of such well-constructed labels are

```
<AbstractText Label="INTRODUCTION AND MATERIAL" NlmCategory="BACKGROUND">
<AbstractText Label="PURPOSE OF INVESTIGATION" NlmCategory="OBJECTIVE">
<AbstractText Label="DESIGN SETTING AND PARTICIPANTS" NlmCategory="METHODS">
<AbstractText Label="MEASUREMENTS AND MAIN FINDINGS" NlmCategory="RESULTS">
<AbstractText Label="CLINICAL SIGNIFICANCE" NlmCategory="CONCLUSIONS">
```

2 Objective

The goal of this project is to assign one of the five canonical NLM categories to structured-abstract[1] labels that have no such NLM category. Labels that are not well constructed fall into three categories, presented next.

2.1 Unmapped Labels (no NlmCategory)

2,259,884 (12.97%) of these label instances (5,155 distinct) have no NlmCategory, e.g.,

```
<AbstractText Label="THE ROLE OF HERBS AND SPICES IN HEALTH">
<AbstractText Label="DATA SOURCES, EXTRACTION, AND SYNTHESIS">
<AbstractText Label="PARTICIPANTS, INTERVENTION, AND MEASURES">
<AbstractText Label="PARTICIPANTS/MATERIALS, SETTING, METHODS">
<AbstractText Label="CONCLUSIONS AND IMPLICATIONS FOR PRACTICE">
```

[1] For this project, we ignore `<OtherAbstract>`s.

2.2 Labels with NlmCategory of "UNASSIGNED"

And 183,838 (1.06%) (11,568 distinct) have an NlmCategory of `UNASSIGNED`, e.g.,

```
<AbstractText Label="INTERVENTIONS IN BARS" NlmCategory="UNASSIGNED">
<AbstractText Label="AN OBSERVATIONAL STUDY" NlmCategory="UNASSIGNED">
<AbstractText Label="THE MODIFICATION OF RISK" NlmCategory="UNASSIGNED">
<AbstractText Label="IMPLICATIONS FOR HEALTH POLICY AND NURSING" NlmCategory="UNASSIGNED">
<AbstractText Label="MORE LESSONS FROM DEVELOPED COUNTRIES FOR IMCI" NlmCategory="UNASSIGNED">
```

2.3 UNLABELLED Labels

Finally, 84,362 (0.48%) have an "UNMAPPED" label, e.g.,

```
<AbstractText Label="UNLABELLED">
```

This document proposes a method of automatically linking unmapped (section 2.1) and UNASSIGNED (section 2.2) labels to one of the five canonical NlmCategories (section 1). We ignore UNLABELLED labels.

3 Methods

We present now three methods for linking most of the unmapped and UNASSIGNED labels presented in section 2.

3.1 Automatic Linking

In 2015, NLM produced a linking of 3,032 distinct labels to one of the five canonical NlmCategories. We will refer to these linkings as the *2015 linkings*; they are available at

```
https://lhncbc.nlm.nih.gov/ii/areas/structured-abstracts/
        downloads/Structured-Abstracts-Labels-102615.txt
```

and look like

```
A CASE REPORT|METHODS|N|20131106
ABBREVIATIONS|BACKGROUND|N|20100629
ACCESS TO DATA|BACKGROUND|N|20131106
ACCME ACCREDITATION|BACKGROUND|N|20131106
ACHIEVEMENTS|RESULTS|Y|20151026
```

Linking unmapped and UNASSIGNED label instances that appear in the 2015 list is automatic; the good news is that

- 1,814 of the 5,155 (35.19%) distinct unmapped labels, and 2,241,658 of the 2,259,884 (99.19%) of the unmapped label instances, and

- 980 of the 11,568 (8.47%) distinct UNASSIGNED labels, and 144,857 of the 183,838 (78.80%) UNASSIGNED label instances

appear in this list, and can therefore be automatically linked.

In the two detailed spreadsheets (one for unmapped, the other for UNASSIGNED labels), automatically linked labels are displayed as `AUTO:RESULTS`, `AUTO:METHODS`, etc.

3.2 Algorithmic Linking by Score

More challenging, however, is the task of linking

- the remaining 3,341 distinct unmapped labels, and 18,226 (0.81%) of the unmapped label instances as well as
- the remaining 10,588 distinct UNASSIGNED labels, and 39,083 (21.26%) of the UNASSIGNED labels instances.

The remainder of this document proposes two algorithmic methods for linking most of the remaining label instances.

In addition to the 2015 linkings (section 3.1), we rely also on the contents of a January 23, 2014 e-mail exchange (included for reference as an appendix) mainly between Lou Knecht and Anna Ripple which established a priority ranking of the five canonical NLM categories (section 1). We assigned numerical ranks to the five categories, from highest to lowest priority:

OBJECTIVE	5
CONCLUSIONS	4
RESULTS	3
METHODS	2
BACKGROUND	1

We take as an example the label `STATEMENT OF SIGNIFICANCE`, which appears 1,783 times in the 2021 Baseline.

3.2.1 Scoring Each Word in Label

We begin by examining the occurrences of each word in the label (excluding PubMed stopwords)[2]) as a token in the 2015 linkings. For example, `STATEMENT` appears 15 times:

```
1 CONCLUDING STATEMENT|CONCLUSIONS
2 CONFLICT-OF-INTERESTSTATEMENT|BACKGROUND
3 CONSENSUS STATEMENT|METHODS
4 IMPLICATION STATEMENT|CONCLUSIONS
5 IMPLICATIONS STATEMENT|CONCLUSIONS
```

[2] `https://pubmed.ncbi.nlm.nih.gov/help/#help-stopwords`

```
 6 PROBLEM STATEMENT|OBJECTIVE
 7 PROBLEM STATEMENT AND BACKGROUND|OBJECTIVE
 8 PROBLEM STATEMENT AND PURPOSE|OBJECTIVE
 9 STATEMENT OF CONCLUSIONS|CONCLUSIONS
10 STATEMENT OF PROBLEM|BACKGROUND
11 STATEMENT OF PROBLEM AND RATIONALE|BACKGROUND
12 STATEMENT OF PROBLEMS|BACKGROUND
13 STATEMENT OF PURPOSE|OBJECTIVE
14 STATEMENT OF THE PROBLEM|BACKGROUND
15 SUMMARY STATEMENT|CONCLUSIONS
```

The category counts for those 15 lines are

OBJECTIVE	4
CONCLUSIONS	5
RESULTS	0
METHODS	1
BACKGROUND	5

We then perform the same calculation on `SIGNIFICANCE`, which appears 36 times in the 2015 list; the category counts for these 36 occurrences are:

OBJECTIVE	0
CONCLUSIONS	32
RESULTS	1
METHODS	0
BACKGROUND	3

3.2.2 Combining Word Scores

We then sum the category counts for each word:

OBJECTIVE	4 + 0	4
CONCLUSIONS	5 + 32	37
RESULTS	0 + 1	1
METHODS	1 + 0	1
BACKGROUND	5 + 3	8

and finally multiply each category's sum by its priority given above:

OBJECTIVE	4 * 5	20
CONCLUSIONS	37 * 4	148
RESULTS	1 * 3	3
METHODS	1 * 2	2
BACKGROUND	8 * 1	8

According to this analysis, CONCLUSIONS has the highest score (148) of the five canonical NLM Categories, and is therefore the winning category. In case of ties (e.g., if OBJECTIVE and CONCLUSIONS both scored 148), the winner is deemed to be the category with the higher priority (OBJECTIVE).

In the two detailed spreadsheets, label instances algorithmically linked by score are displayed as `SCORE (159):RESULTS`, `SCORE (296):METHODS`; the number in parentheses is the maximum score across the five canonical NLM categories.

This algorithm successfully linked

- 2,800 of the remaining 3,341 (83.80%) distinct unmapped labels, and 16,998 of the 18,226 (93.26%) remaining unmapped label instances (section 2.1) that could not be automatically linked (section 3.1), as well as
- 8,675 of the remaining 10,588 (81.93%) distinct UNASSIGNED labels, and 28,376 of the 39,083 (72.60%) remaining UNASSIGNED label instances (section 2.1) that could not be automatically linked (section 3.1).

After applying both automatic linking and algorithmic linking by score, we have successfully linked

- 4,724 of all the 5,155 (91.64%) distinct unmapped labels and 2,258,680 of all the 2,259,884 (99.95%) unmapped label instances (section 2.2) that could not be automatically linked (section 3.1), and
- 9,950 of all the 11,568 (86.01%) distinct UNASSIGNED labels, and 173,978 of the 183,838 (94.64%) UNASSIGNED label instances (section 2.2) that could not be automatically linked (section 3.1).

3.3 Algorithmic Linking by Minimum Edit Distance

The second algorithmic linking computes, for each as-yet-unlinked label, the minimum edit distance[3] to all label instances that had been successfully been linked either automatically or algorithmically by score. This strategy was especially fruitful for linking

1. misspellings, e.g., `abstarct`, `backgroud`, `conslusions`, `ntroduction`, `objetive`; `pupose`;
2. non-English (principally Spanish and Portuguese) terms, e.g., `caso clinico`, `intervenciones`, `introduccion`, `resultados`, `conclusao`, `introducao`;
3. plural forms whose singular form *does* appear in the 2015 list, e.g., `concepts`, `contributions`, `diagnoses`, `responses`.

Inspection of results revealed that minimum edit distances of

- less than 4 led to solid linkings via a similar linked label, e.g.,

[3] See e.g., `https://en.wikipedia.org/wiki/Edit_distance`, `https://observablehq.com/@stwind/minimum-edit-distance`, and `https://web.stanford.edu/class/cs124/lec/med.pdf`.

  - `diagnoses` ⇒ `DISTANCE (1):diagnosis:METHODS`
  - `fundamento` ⇒ `DISTANCE (3):fundamentals:OBJECTIVE`

- 4 or 5 led to questionable linkings requiring human review; sample output for the label `drugs` is

```
REVIEW (4):aims/SCORE (355):OBJECTIVE;
           goals/SCORE (80):OBJECTIVE;
           focus/SCORE (10):OBJECTIVE;
```

  4 is the minimum edit distance for the label `drugs`; 355, 80, and 10 are the results of the score algorithm (Section 3.2) for `aims`, `goals`, and `focus`, respectively/

- greater than 5 were essentially noise.

After adding minimum-edit distance linking (distance < 4) in addition to automatic linking and algorithmic linking by score, we are pleased to report that we have linked

- 4,967 of all the 5,155 (96.35%) distinct unmapped labels and 2,259,488 of all the 2,259,884 (99.98%) unmapped label instances, as well as
- 10,520 of all the 11,568 (90.94%) distinct UNASSIGNED labels, and 181,674 of the 183,838 (98.82%) UNASSIGNED label instances.

These numbers will probably receive a further slight boost after human review of the questionable label linkings.

Remaining unlinked labels fall into several categories:

- Possible errors or highly technical or specific terms, e.g., "`, ffw`", "`pbh-lci`", "`bw, adg, f`";
- Foreign terms, e.g., `ergebnisse`, `einleitung`, `fortolkning`;
- Perfectly normal labels that unfortunately have no similarity to any previously linked label, e.g., `drugs and falls`, `laser parameter`, `pleural lesions`, `lymphatic tissue`, `cellular viability`, etc.
- Exceptionally long labels, e.g., `gait mainly depends on the relationship between posture balance and movement`, `linear versus sigmoid relationship between blood pressure fall and drug concentration`, `polymerization shrinkage stress and stress reduction possibilities`, etc.

4 Results

The table below summarizes all the above results.

	TOTAL			
	distinct labels		label instances	
	count	%age	count	%age
unmapped	5,155	100.00	2,259,884	100.00
UNASSIGNED	11,568	100.00	183,838	100.00

	Linked after AUTO			
	distinct labels		label instances	
	count	%age	count	%age
unmapped	1,814	35.19	2,241,658	99.19
UNASSIGNED	980	8.47	144,857	78.80

	Linked after AUTO and SCORE			
	distinct labels		label instances	
	count	%age	count	%age
unmapped	4,724	91.64	2,258,680	99.95
UNASSIGNED	9,950	86.01	173,978	94.64

	Linked after AUTO, SCORE, and DISTANCE			
	distinct labels		label instances	
	count	%age	count	%age
unmapped	4,967	96.35	2,259,488	99.98
UNASSIGNED	10,520	90.94	181,674	98.82

5 Conclusion

Notice that in the final chart above (Linked after AUTO, SCORE, and DISTANCE), in both the unmapped and UNASSIGNED rows, the percentage of distinct labels linked (**96.35**% and **90.94**%) is lower than the percentage of label instances linked (**99.98**% and **98.82**%). This difference points to a long tail of low-frequency labels that could not be linked to one of the five canonical NlmCategories. Indeed, labels that were successfully linked have an average frequency of 45,490 (unmapped) and 1,727 (UNASSSIGNED), whereas labels that could not be linked have an average frequency of 2 (for both unmapped and UNASSIGNED).

Automatic and algorithmic linkings have served to link a vast majority of unmapped and UNASSIGNED labels. We hope that these results might lead to the linking of most Structured-Abstract labels in the Baseline.

Appendix

In 2014, NLM subject-matter experts ranked the 5 canonical metadata NlmCategories (BACKGROUND, OBJECTIVE, METHODS, RESULTS, CONCLUSIONS) by order of importance for mapping structured-abstract labels and information-retrieval experimentation. The rankings are as follows:

1. OBJECTIVE
2. CONCLUSIONS
3. RESULTS
4. METHODS
5. BACKGROUND

OBJECTIVE is ranked first because of the overall contextual importance of a study's purpose. CONCLUSIONS is ranked ahead of RESULTS because in the majority of cases, CONCLUSIONS does limit itself to the major results; some branch off into the further-research arena, but not enough to skew the main thrust of an article. Moreover, CONCLUSIONS is supported by results/discussion in the article. METHODS is ranked fourth because of the tendency to find non-semantic-type concepts indicative of patient population characteristics, experimental animals, and publication types. BACKGROUND is last because it usually discusses peripheral information related to the study.